ARTICLE TYPE

Recurrent and Convolution Neural Networks for Sequential Multispectral Optoacoustic Tomography (MSOT) Imaging

Aniwat Juhong [1,9] | Bo Li [1,9] | Yifan Liu [1,9] | Cheng-You Yao [2,9] | Chia-Wei Yang [3,9] | Dalen W. Agnew [4] | Yu Leo Lei [5] | Gary D. Luker [6] | Harvey Bumpers[7] | Xuefei Huang [2,3,9] | Wibool Piyawattametha [8,9] | Zhen Qiu* [1,2,9]

1 Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, USA

2 Department of Biomedical Engineering, Michigan State University, East Lansing, Michigan, USA

3 Department of Chemistry, Michigan State University, East Lansing, Michigan, USA

4 Department of Pathobiology and Diagnostic Investigation, College of Veterinary Medicine, Michigan State University, East Lansing, Michigan, USA

5 Department of Periodontics and Oral Medicine, University of Michigan, Ann Arbor, Michigan, USA

6 Department of Radiology, Microbiology and Immunology, and Biomedical Engineering University of Michigan, Ann Arbor, Michigan, USA

7 Department of Surgery, Michigan State University, East Lansing, Michigan, USA

8 Department of Biomedical Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand

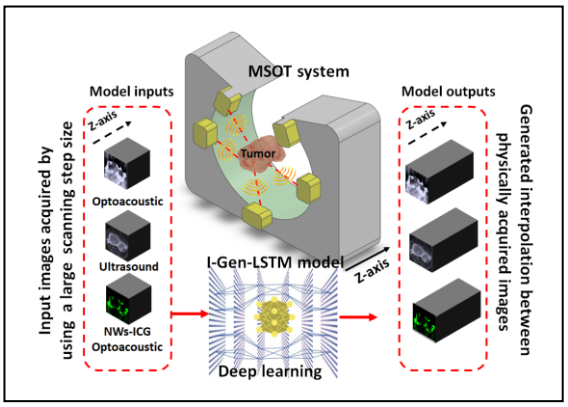
9 Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, USA

*** Correspondence**

Zhen Qiu, Department of Biomedical Engineering, Michigan State University, East Lansing, Michigan, USA

Email: qiuzhen@msu.edu

Volumetric optoacoustic imaging is a beneficial technique for diagnosing and analyzing biological samples since it provides meticulous details in anatomy and physiology. However, acquiring high through-plane resolution volumetric images is time-consuming, requiring a precise motorized stage to move samples under the optoacoustic system along the z-axis. Here, we propose deep learning based on hybrid recurrent and convolution neural networks to generate sequential cross-sectional optoacoustic images. A multispectral optoacoustic tomography (MSOT) system was utilized to acquire the dataset from breast tumors for training our deep learning model. This system can simultaneously acquire the sequential images (cross-sectional images) of MSOT and ultrasound. Furthermore, it provides a spectral unmixing algorithm applied to the MSOT images for extracting the sequential images of a specific exogenous contrast agent. This study used ICG-conjugated superparamagnetic iron oxide nanoworms particles (NWs-ICG) as the contrast agent. Our deep learning model applies to all three modalities (multispectral optoacoustic imaging at a specific wavelength, ultrasound, and NWs-ICG optoacoustic imaging). The generated 2D sequential images were compared to the ground truth 2D sequential images acquired using a small step size. The results of these three modalities can achieve excellent image quality where the average of peak-signal-to-noise ratio and summation absolute errors between the ground truths and the generated images is over 75 dB and less than 2,000. Instead of acquiring seven images with a step size of 0.1 mm, we can receive two images with a step size of 0.6 mm as input images for the proposed deep learning model. The deep learning model can generate or interpolate other five images with the step size of 0.1 mm between these two input images meaning we can save acquisition time by approximately 71%.



KEYWORDS

convolution neural network, recurrent neural network, multispectral optoacoustic tomography, volumetric imaging

1 | INTRODUCTION

Multispectral Optoacoustic Tomography (MSOT) is an *in vivo* optical imaging modality for molecular, anatomical, and functional imaging Fields [1, 2]. The principle of MSOT is based on the optoacoustic effect, i.e., a molecule is excited by an ultra-short laser pulse, which can penetrate through tissue several centimeters [3, 4], resulting in thermoelastic expansion surrounding the molecule that generates a photoacoustic wave [5]. The ultrasound traducer is then used to detect this wave as an ultrasound signal. The difference of absorption contrast of tissue in single wavelength images is employed to reconstruct anatomical images. Using multiple wavelengths to excite the tissue, we can obtain multispectral images from intrinsic and extrinsic signals. A laser between 680 nm and 980 nm is the predominant source for intrinsic signals such as deoxygenated

hemoglobin, oxygenated hemoglobin, melanin, myoglobin, bilirubin, fat, etc. Extrinsic signals do not usually occur in cells, tissue, or animals. Agents that can absorb in the near-infrared (NIR) range such as indocyanine green, fluorescence proteins, nanoparticles, etc., can increase the optoacoustic signal (extrinsic signal). Thus, they can be distinguished from intrinsic tissue background signals by using effective spectral unmixing algorithms such as linear regression, guided independent comment (ICA), and principal component analysis (PCA) [6, 7]. MSOT is widely used for several studies such as cancer research [8-12], drug development [13, 14], and nanoparticle [15-18]. However, using multiwavelength excitation to scan the sample is time-consuming, especially cross-sectional scanning for 3D image reconstruction. Imaging needs to sweep all the wavelengths with every single scanning position. For *in vivo* experiments, this might lead to image degradation from motion artifacts and potential lethality from prolonged anesthesia. In recent years, deep learning-based approaches have played a vital role in optoacoustic imaging, and they have been widely used in several applications such as image classification, segmentation [19-23], quantitative photoacoustic imaging [24-28], image enhancement [29-33], etc. One main advantage of deep learning for those applications is that it depends less on hardware modifications. In addition, most of those deep learning techniques were designed to use a single 2D image as their input and apply convolution architectures for feature extraction. For instance, deep learning for automatic segmentation of optoacoustic ultrasound (OPUS) images [34] used the U-net architecture [35] to perform the image segmentation. U-net is a well-known convolution neural network (CNN) architecture for image segmentation, particularly biomedical images [36-39].

Nevertheless, there are no techniques based on deep learning to reduce the acquisition time of cross-sectional scanning for 3D photoacoustic imaging. Herein, we propose the hybrid architecture of convolution neural network (CNN) and recurrent neural network (RNN) for generating sequential optoacoustic, unmixed optoacoustic of a specific contrast agent, and ultrasound images to extend the stack of cross-sectional images and reduce acquisition time by approximately 71%. This hybrid architecture is called Inception Generator Long Sort-Term Memory (I-Gen-LSTM). The Inception Generator is a CNN model designed based on the Inception U-net architecture. Inception is a convolution layer [40] that convolves the input in parallel with different kernel sizes extracting more features than a simple convolution layer. RNN is a robust and effective approach for sequential problems. It is a feed-forward neural network with internal memory and performs the same function for every data input. In addition, the output of the current input depends upon the previous output. However, the original RNN has drawbacks regarding exploding and vanishing gradients from backpropagation to update weights, particularly long sequential inputs. Long Short-Term Memory (LSTM) networks [41] are improved RNN networks capable of learning long-term dependencies by adding a forget gate, input gate, and output gate. Therefore, we leverage Inception Generator and LSTM networks to generate sequential images. Our results demonstrate that the I-Gen-LSTM model is a versatile method that can generate not only sequential optoacoustic images but also sequential unmixed optoacoustic and ultrasound images.

2 | EXPERIMENTAL

2.1 | Data acquisition

A commercial multispectral optoacoustic tomography (MSOT) system (inVision 512-echo, iThera Medical GmbH, Munich, Germany) was used to acquire the data for training the I-Gen-LSTM model. The MSOT system has a 270-degree ultrasound transducer tomographic array, which can acquire signals from multiple angles around an object. This tomographic array enables the system for imaging complex shapes since it can capture 2-dimensional signals in the imaging plane. Figure 1(a) shows the detection and illumination geometry in the imaging chamber of the MSOT system. In addition, this system provides a tunable laser with a range of 660-1,300 nm, which is particularly suitable for most biological samples. The excitation pulse laser is used to illuminate the sample. The sample absorbs this pulse and converts it to heat, which results in a transient thermoelastic expansion that generates an acoustic wave. The ultrasound transducer is then used to detect this acoustic wave, and the back-projection algorithm [42] is applied to the detected optoacoustic wave to reconstruct the images. For the dataset preparation, transgenic mice [43] with breast tumors were intravenously injected with indocyanine green (ICG)-conjugated superparamagnetic iron oxide nanoworms (NWs-ICG) [44], which accumulate in tumors longer than pure ICG through the enhanced permeability and retention (EPR) effect [45]. Twenty-four hours after injection, the mice were euthanized and the tumors were removed and dissected for this study. All procedures performed on animals were approved by the University's Institutional Animal Care & Use Committee and were within the guidelines of humane care of laboratory animals. To acquire images of the tumors, 4 mg of agarose powder was dissolved in 40 mL of warm deionized water. The breast tumor was put in this dissolved agarose solution, allowing approximately 15 minutes for the solution to solidify. The hardened agarose with the tumor inside shown in Figure 1(b), was grasped by the holder and then scanned by the inVision MSOT system with the excitation pulse at wavelengths from 800 nm to 1000 nm (a comprehensive range of the NWs-ICG study). Since the inVision MSOT system can provide corresponding ultrasound images, NWs-ICG optoacoustic images obtained through linear spectral unmixing algorithm [46], and each single-wavelength optoacoustic image, these three imaging modalities were simultaneously acquired in every scanning

position. Figure 1(d1-d4) shows the ultrasound images of the breast tumor with different scanning positions, Figure 1(e1-e4) shows the corresponding NWs-ICG optoacoustic images reconstructed from multispectral optoacoustic imaging with the excitation pulse at wavelengths from 800 to 1,000 nm by using the multispectral unmixing algorithm; Figure 1(f1-f4) shows the corresponding single-wave optoacoustic image at 800 nm excitation; and Figure 1(g1-g4) shows the corresponding overlaid images of these three imaging modalities.

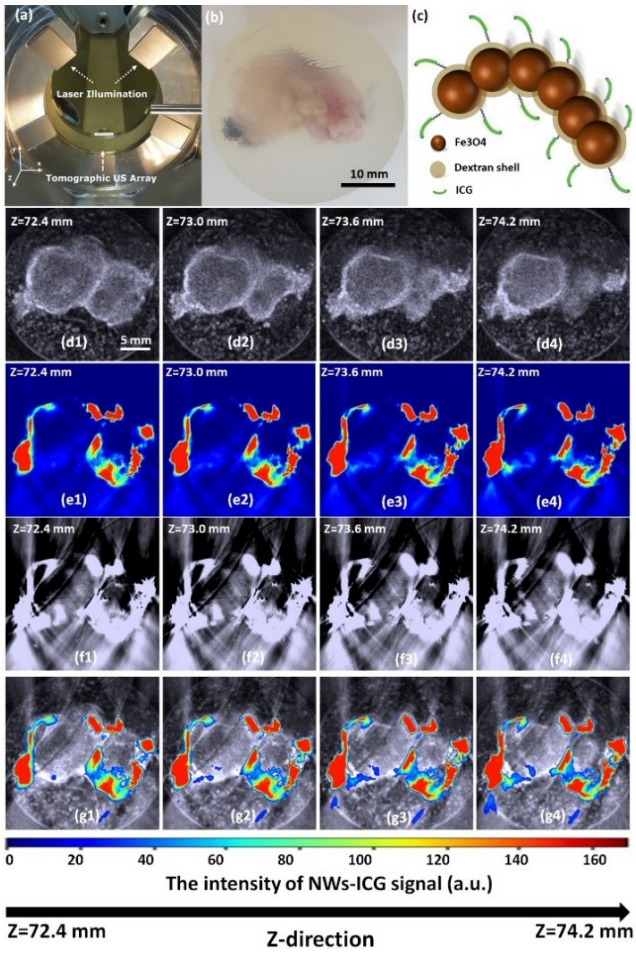


Figure 1. Ultrasound, NWs-ICG optoacoustic obtained through multispectral unmixing, and optoacoustic at 800 nm excitation imaging of an ex vivo breast tumor from a mouse intravenously injected with NWs-ICG. (a) The detection and illumination geometry in the imaging chamber of the MSOT system. (b) The breast tumor is embedded in agarose. (c) NWs-ICG structure. (d1-d4) Ultrasound images of the breast tumor with different step sizes. (e1-e4) The corresponding NWs-ICG optoacoustic images were obtained through multispectral unmixing. (f1-f4) The corresponding single-wavelength (λ_{ex} = 800 nm) optoacoustic images. (g1-g4) with an overlay of the ultrasound, the NWs-ICG optoacoustic(colormap), and the single-wavelength optoacoustic images.

2.2 | I-Gen-LSTM and discriminator models

The I-Gen-LSTM model comprises three main neural networks depicted in Figure 2(a-c). The first neural network is the Inception encoder & decoder network based on Inception U-net architecture. The original U-net architect employs simple convolution blocks with the skip connection of encoders and decoders at the same dimension helping the model to circumvent the vanishing and exploding gradients problems. However, the simple convolution blocks might be insufficient to extract all crucial information comprehensively. Inception architecture is one of the effective CNNs architectures since it applies a wide range of kernel sizes to extract global and local features. A large and a small kernel size are tailored to extract information distributed globally and locally, respectively. With this attribute, the encoder & decoder network was designed using Inception U-net as its backbone as shown in Figure 2(a), for improving the model capability. This network takes two 2D images, acquired from an arbitrary consecutive position with a step size of 0.6 mm, as its inputs (input 1 and input 2, as shown in Figure 2(a)). The encoder shown on the left side of Figure 2(a) generates encoder outputs (E1n -E5n, where n is the input image number, i.e., 1 and 2). Inception architecture in the encoder with three different kernel sizes (1x1, 3x3, and 5x5) assembled as the parallel filters are used to extract features from the tensors followed by a rectified linear unit (ReLU) and a 2x2 max pooling with the stride of 2 steps for downsampling, respectively. Similarly, Inception architecture is also used in the decoder blocks. The encoder blocks are used to generate decoder outputs (D1n-D5n, where n is the input image number, i.e., 1 and 2) as shown in the right side of Figure 2(a) followed by a feature map upsampling, a 2x2 up-convolution (halving the number of feature channels), and a corresponding concatenation from the encoder part.

The second neural network is the convolutional LSTM network (ConvLSTM) [47], a recurrent neural network for spatio-temporal prediction. It has a convolutional structure in both the input-to-state and state-to-state transitions as shown in the bottom right of Figure 2(b). In other words, internal matrix multiplications are exchanged with convolution operations. Consequently, the data flowing through the ConvLSTM cells keeps the input dimension instead of being a 1D vector with features. The main equations of ConvLSTM are expressed in Equations (1-5) below, where '$*$' and '$\circ$' represent the convolution operator and the Hadamard product (element-wise matrix multiplication), respectively. All variables in Equations (1-5) were shown in the "ConvLSTM block" in Figure 2(b).

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} \circ C_{t-1} + b_i) \quad (1)$$

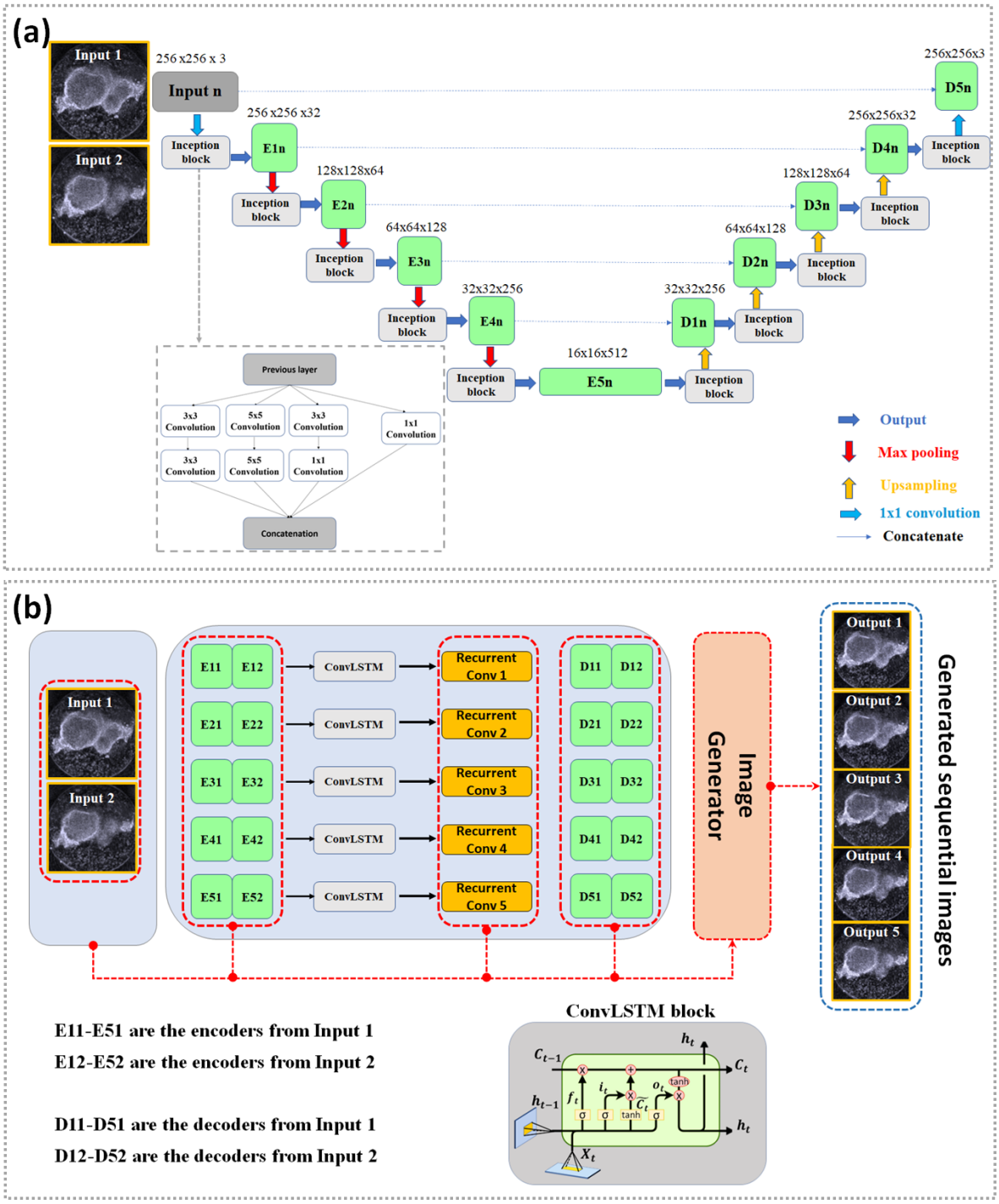
$$f_t = \sigma(W_{xf} * X_t + W_{ht} * H_{t-1} + W_{ct} \circ C_{t-1} + b_f) \quad (2)$$

$$c_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_c) \quad (4)$$

$$H_t = o_t \circ \tanh(C_t) \quad (5).$$

The ConvLSTM takes the outputs of the Inception encoder from both input images (E11-E51 and E12-E52) as its inputs to generate five sequential blocks (Recurrent Conv1 to Recurrent Conv5) as shown in Figure 2(b). Recurrent Conv 1, 2, 3, 4, and 5 have dimensions of (5x128x128x512), (5x64x64x512), (5x32x32x512), (5x16x16x512), and (5x8x8x512), respectively. The first dimension represents the number of output images (five sequential output images).



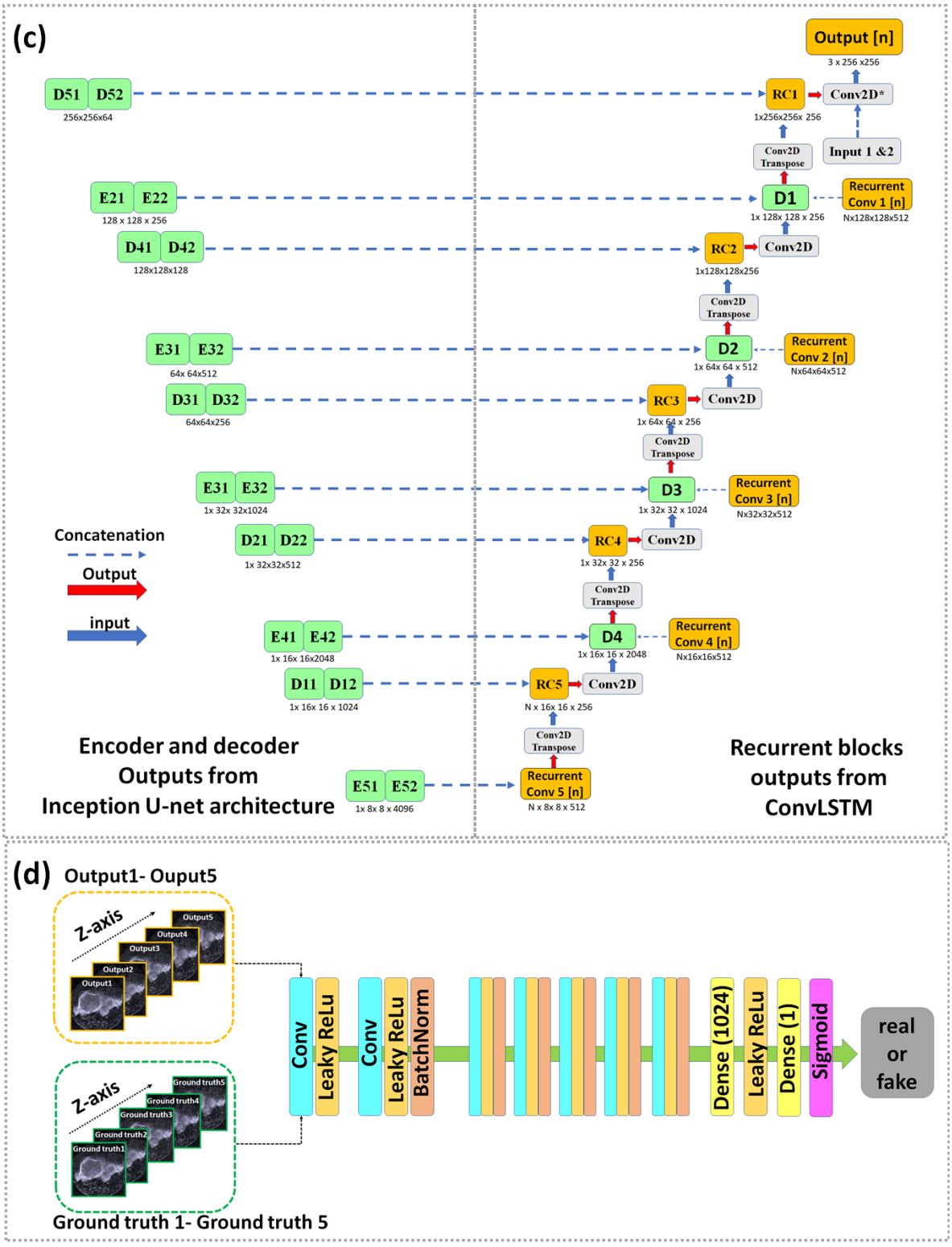


Figure 2. I-Gen-LSTM and discriminator architectures. (a) Inception encoder and decoder network were applied to both images (input1 and input2). (b) ConvLSTM network for generating the sequential blocks (Recurrent Conv 1-5) fed to the sequential image generator network for reconstructing the sequential output images. (c) The sequential image generator network. (d) The discriminator network.

Lastly, it is the sequential image generator network inspired by U-net architecture. The model takes Recurrent Conv 1-5, two input images, encoder outputs (E11-E51 and E12-E52), and decoder outputs (D11-D41 and D12-D42) to reconstruct five sequential images of different scanning positions as shown in Figure 2(c). The left side of Figure 2(c) shows the concatenated encoder and decoder outputs generated by the Inception encoder &decoder (Figure 2(a)). The right side of Figure 2(c) shows Conv2D transpose and Conv2D operations for the Recurrent Conv 1-5 generated by the ConvLSTM blocks (Figure 2(b)) and the concatenated encoder & decoder outputs. All Conv2D transpose, Conv2D blocks utilize ReLU as their activation function except the last Conv2D* that applies hyperbolic tangent or tanh as its activation function. Indeed, the Recurrent Conv blocks regulate the gradual change in the sequential output images. In short, the I-Gen-LSTM model takes two images acquired by consecutive positions with 0.6 mm steps size and generates the five sequential images between these two images with gradual change following the scanning positions (step sizes of 0.1 – 0.5 mm). The ground truth images acquired using a small step size (0.1-0.5 mm) were used to determine the loss value from these five generated images. The loss functions will be elucidated in section 2.3.

The discriminator network shown in Figure 2(d) is a simple convolution network designed to evaluate the similarity between the ground truths and generated images. The model comprises eight convolutional layers and two fully connected layers. After each convolution block, a batch normalization layer is used, followed by an activation function named the Leaky ReLU function (α=0.2). The number of 3x3 filter kernels increases by a factor of 2 from 64 (the first layer) to 512 (the eighth layer) kernels. The last two layers are dense layers working as a classification block, predicting the probability of an image being either real or fake. To train the I-Gen-LSTM model, we assemble the models as a generative adversarial network (GAN) [48] shown in Figure 3 below.

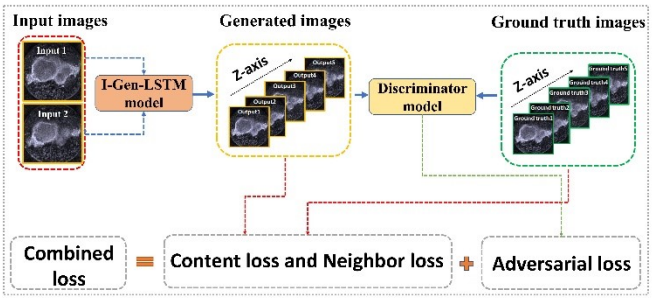


Figure 3. GAN with the combination of three loss functions (the content loss, the neighbor loss, and the adversarial loss functions) for training the I-Gen-LSTM model.

2.3 | Loss functions

To optimize the I-Gen-LSTM model, we designed custom-made loss functions, namely the content loss (VGG19 loss, I_{VGG}^{SS}) [49], adversarial loss (Discriminator loss, I_{Gen}^{SS}), and neighbor loss (I_{N}^{SS}) as shown in Equation (6). Where C_{w1}, C_{w2}, and C_{w3} are the hyper-parameters set as 0.7, 0.1, and 0.2, respectively.

$$I^{SS} = C_{w1} I_{VGG}^{SS} + C_{w2} I_{Gen}^{SS} + C_{w3} I_{N}^{SS} \quad (6)$$

The content loss or VGG loss (I_{VGG}^{SS}), which is defined as the Euclidean distance between the feature map of the generated image ($G_{\theta G}(I^{LS})$) and the ground truth (I^{SS}), can extract high dimensional features helping the model to generate the image with perceptually satisfying solutions without excessively smooth textures. The I_{VGG}^{SS} loss is based on the ReLU activation layers of the pre-train 19-layer VGG network and it can be calculated following Equation (7) as shown as

$$I_{VGG}^{SR} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\emptyset_{i,j}(I^{SS})_{x,y} - \emptyset_{i,j}(G_{\theta G}(I^{LS})_{x,y}) \right)^2 \quad (7)$$

where $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network. The features map ($\emptyset_{i,j}$) can be obtained by the j-th convolution before the i^{th} maxpooling layer within the VGG19 network.

Moreover, the adversarial loss (I_{Gen}^{SS}) is also employed to distinguish the similarity of the two images. It is defined as the probabilities, varying from 0 to 1, which are the result of the discriminator model ($D_{\theta_D}(G_{\theta_G}(I^{LS}))$) as shown in Equation (8). Where I^{LS} is the input images, G_{θ_G} is the generator model, and D_{θ_D} is the discriminator model.

$$I_{Gen}^{SS} = \sum_{n=1}^{N} -log D_{\theta_D}(G_{\theta_G}(I^{LS})) \quad (8)$$

Apart from using the content and adversarial losses, the neighbor loss is also applied to optimize the model. Since the I-Gen-LSTM model generates sequential images, the neighbor loss is essential to regulate the change of each generated image in the sequence. The concept of the neighbor loss function is to differentiate between the current generated image and the neighbor images in the same sequence as expressed in Equation (9) below as

$$I_{N}^{SS} = \sum_{n=1}^{N} (mse(I_n, I_{n-1}) + mse(I_n, I_{n+1})) \quad (9)$$

The custom-made loss function effectively leverages the combination of these three loss functions to train the I-Gen-LSTM model that can generate high-quality sequential images.

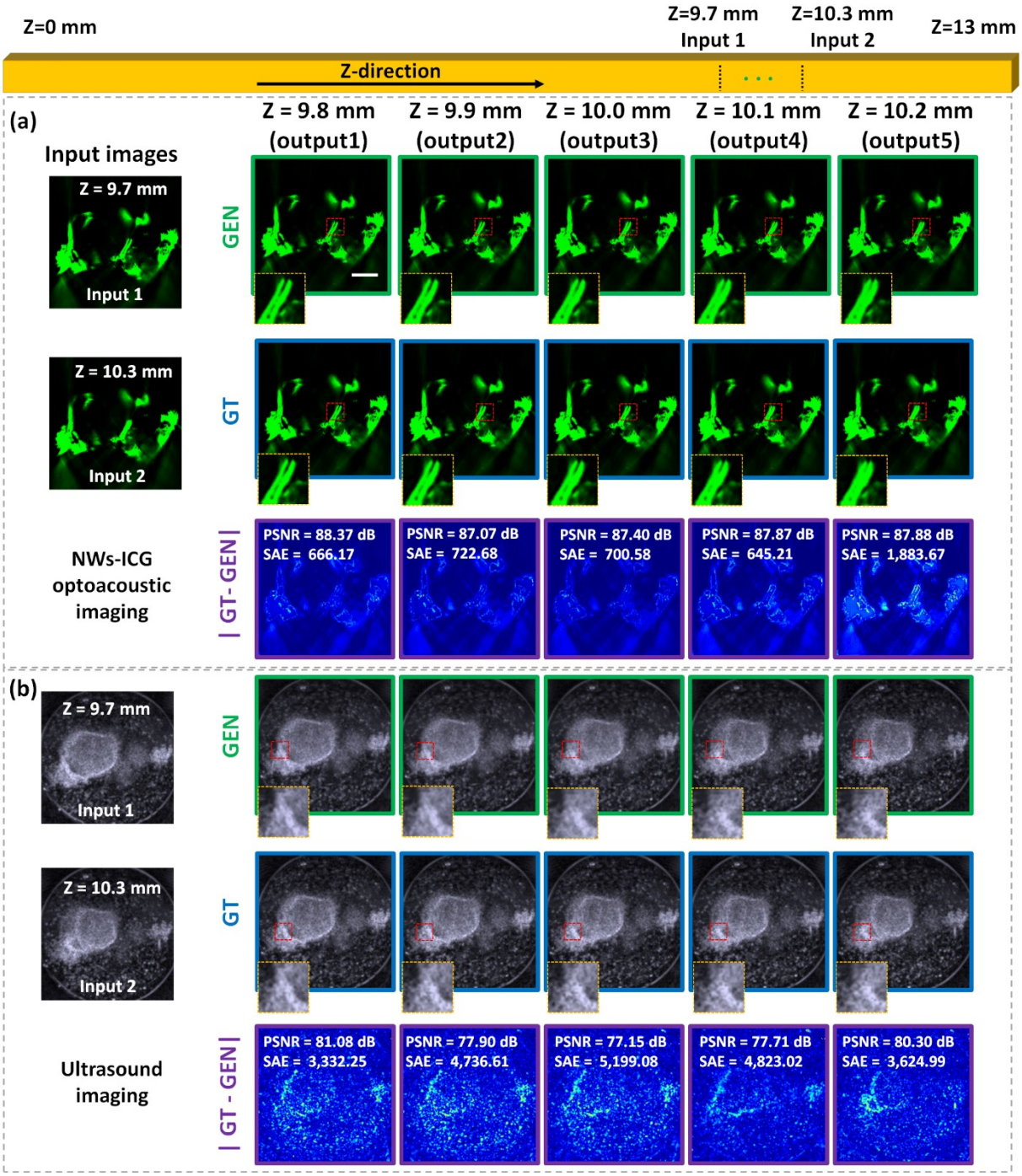
2.4 | I-Gen-LSTM model for Volumetric Imaging

To collect the database for training the model, 16 breast tumors from mice intravenously injected with NWs-ICG were acquired by the MSOT system. The data from these tumors were allocated for training (11 tumors), validation (3 tumors), and testing (2 tumors) datasets. The training time on Google Colaboratory (CoLab) Pro is approximately 40 hours. After initializing and importing the model, the I-Gen-LSTM can generate five sequential images by taking less than 1 second for the five output images on a personal computer (PC) with 11th Gen Intel core i7-11700k CPU, 16 GB RAM, and NVIDIA RTX 3090 graphic card.

3 | Results and Discussion

3.1 | Sequential NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) image reconstruction.

The breast tumor dissected from an NWs-ICG-injected mouse was scanned under the MSOT system. Figure 4 shows the generated sequential images generated by the I-Gen-LSTM model. Two input images of each modality, acquired from consecutive stage positions with a step size of 0.6 mm, are used as the inputs for the I-Gen-LSTM model.



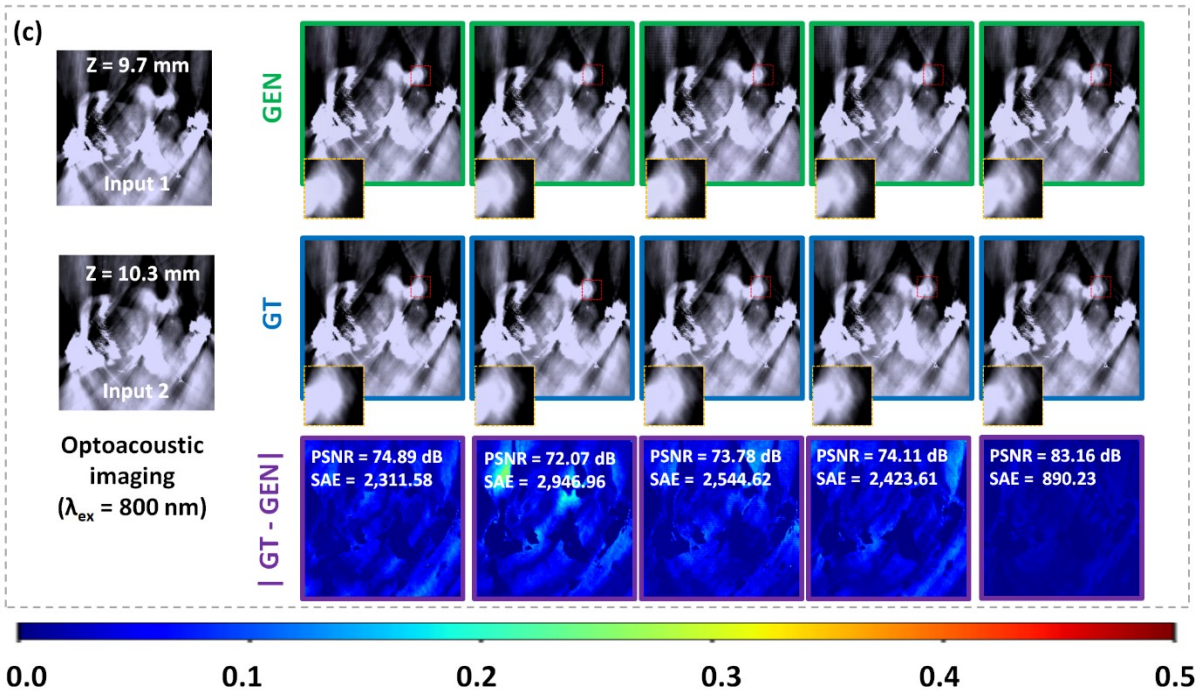


Figure 4. Results of sequential image reconstruction generated by the I-Gen-LSTM model. The two input images for each modality simultaneously acquired with a step size of 0.6 mm were fed into the I-Gen-LSTM model. The green, blue, and violet boxes show generated images (GEN), ground truth (GT), and the absolute error between GEN and GT images (|GT-GEN|) represented as color map images. The red-dashed boxes show the local features fairly change along the z-scanning position and the yellow-dashed boxes are the corresponding enlarged images of the red-dashed boxes. The scale bar is 5 mm. (a) NWs-ICG optoacoustic sequential image reconstruction result. (b) Ultrasound sequential image reconstruction result. (c) Single-wavelength optoacoustic (λ_{ex} = 800 nm) reconstruction result.

Here, we demonstrate a z-scanning range from 9.7 mm-10.3 mm with a step size of 0.1 mm as a representative. The red-dashed boxes in Figure 4 show local features, which are fairly changing along the z-scanning position and are somewhat straightforward to observe. The orange-dashed boxes are the corresponding enlarged images of the red-dashed boxes. Figure 4(a) shows the sequential image reconstruction result of NWs-ICG optoacoustic imaging, Figure 4(b) shows the result of ultrasound imaging, and Figure 4(c) shows the result of single-wavelength optoacoustic (λ_{ex} = 800 nm) imaging. The average Peak-signal-to-noise ratio (PSNR) dB/ the average summation of absolute errors (SAE) between the ground truths (GT) and generated images (GEN) for this scanning range of NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) imaging are 87.72 dB/923.66 ,78.83 dB/4,323.19, 75.60 dB/2,223.40, respectively.

3.2 | Three-dimensional reconstruction of the stack 2D NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images

Since the MSOT system and our deep learning model provide the stack of multiple cross-sectional images for NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images, we can use these images to reconstruct three-dimensional (3D) images by using Amira (Mercury Computer system, Berlin, Germany) software. Figure 5 shows the 3D reconstruction results of the ground truth and the generated images. Figure 5(a) demonstrates the 3D reconstruction of generated images from the I-Gen-LSTM model and Figure 5(b) shows the reconstruction of the ground truths acquired by mechanical scanning. After finished the experiment, the tumor was removed from the agarose and sent to the histopathology lab (MSU-IHPL Research facility) to prepare a Hematoxylin-and-Eosin (H&E) stained breast tumor slide shown in Figure 5(c).

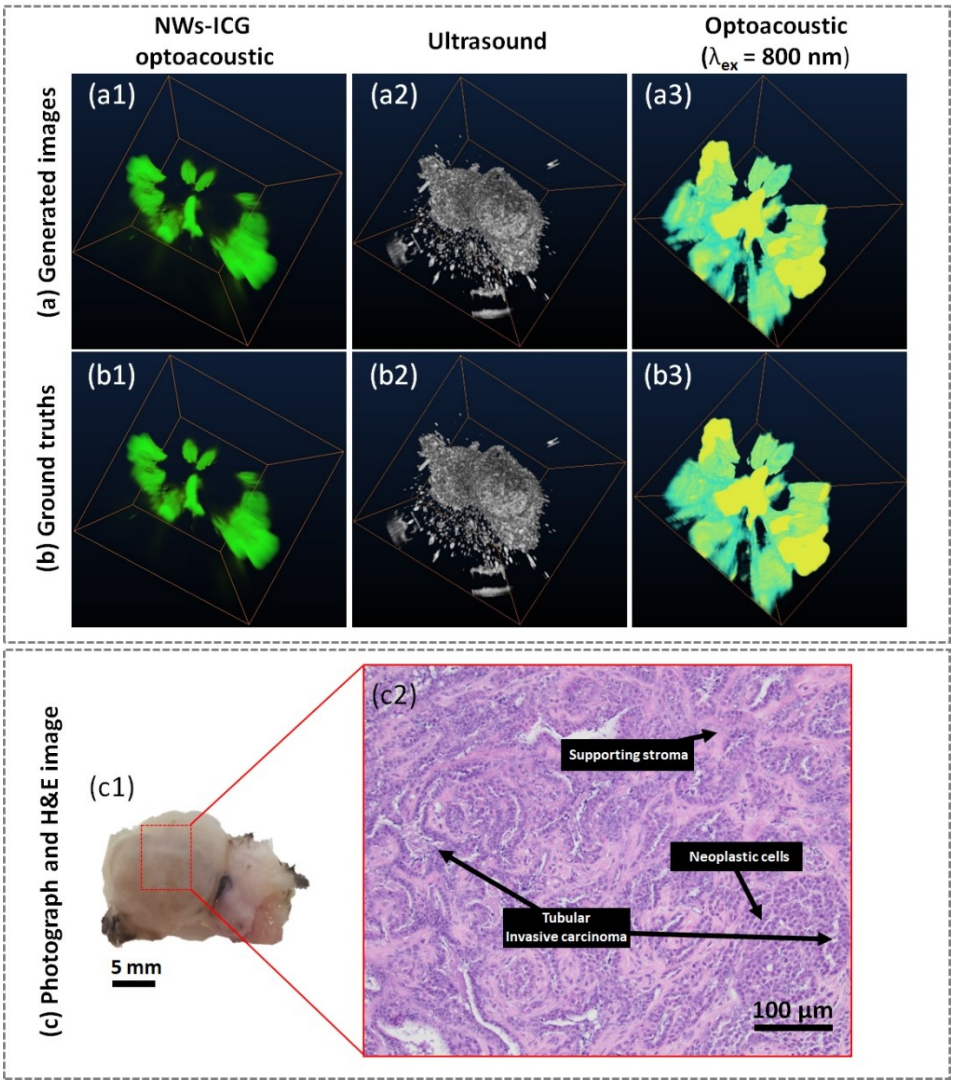


Figure 5. 3D image reconstruction of the breast tumor using cross-sectional NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) stacked images. (a) The 3D reconstruction result of the NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images generated by the I-Gen-LSTM model with a step size of 0.1 mm. (b) The 3D reconstruction result acquired by mechanical scanning with a step size of 0.1 mm. (c) The photograph of the corresponding tumor and its H&E slide image.

3.3 | Evaluations

The NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images from two tumors not used for training the model were utilized for the model evaluation. Each tumor was scanned with a step size of 0.1 mm. Every two-image (with a 0.6 mm scanning step in between) was assigned as the input for the I-Gen-LSTM model to generate five sequential images with a step size of 0.1 mm. Here, the model was evaluated using four quantitative metrics: the average PSNR, SAE (GEN, GT), SAE $(Input_1, GT)$,and SAE $(Input_2, GT)$. They were applied to the testing dataset acquired from the tumors for all scanning positions.

A large PSNR and a small SAE (GEN, GT) imply high-quality generated images. Indeed, if the SAE (GEN, GT) can perform better than SAE (Input1-GT) and SAE (Input2-GT), it also means that the model can effectively generate sequential images. All average evaluation metrics can be calculated following Equations (10-12).

$$\text{Average PSNR} = \frac{\sum_j^N \sum_i^5 PSNR_j(GEN_i, GT_i)}{5 \times N} \quad (10)$$

$$\text{Average SAE (GEN, GT)} = \frac{\sum_j^N \sum_i^5 SAE_j(GEN_i, GT_i)}{5 \times N} \quad (11)$$

$$\text{Average SAE } (Input_k, \text{GT}) = \frac{\sum_j^N \sum_i^5 SAE_j(Input_k, GT_i)}{5 \times N} \quad (12)$$

Where,

N is the number of scanning positions with a step size of 0.6 mm,

GEN_i is the generated image at "i" scanning position in between two input images (acquired with a step size of 0.6 mm),

GT_i is the corresponding ground truth,

$Input_k$ images are the two input images (k=1 and 2) acquired from arbitrary consecutive positions with a step of 0.6 mm.

Figure 6 shows the representative result from one of the evaluated tumors as the graph of the average PSNR and SAE (GEN, GT) vs. scanning positions. Table 1 shows the average evaluation metrics of the generated sequential NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex}= 800 nm) images for all testing datasets. Overall, the average PSNR and SAE between generated images and ground truths of all modalities are greater than 75 dB and less than 2,000, respectively.

This indicates that the I-Gen-LSTM model can generate sequential images with promising results. To comprehensively evaluate the model performance, we also compared SAE (GEN, GT) to SAE ($Input_1, GT$) and SAE($Input_2, GT$) as the baseline for comparison. The average SAE (GEN, GT) of optoacoustic (λ= 800 nm) and ultrasound imaging performs better than the average SAE($Input_1, GT$) and SAE($Input_2, GT$), but the NWs-ICG optoacoustic imaging does not (the average SAE (GEN, GT) is slightly higher than the average of SAE($Input_1, GT$) and SAE($Input_2, GT$)) due to the tiny changing features in the sequential NWs-ICG optoacoustic imaging and the limited number of the training dataset. Although the overall result is favorable and encouraging, the deep learning model could be improved in future work. We will use a larger dataset with a larger image size to train the deep learning model so that the convolution/LSTM blocks can efficiently capture more sequential features, especially in a tiny changing feature modality such as NWs-ICG optoacoustic imaging.

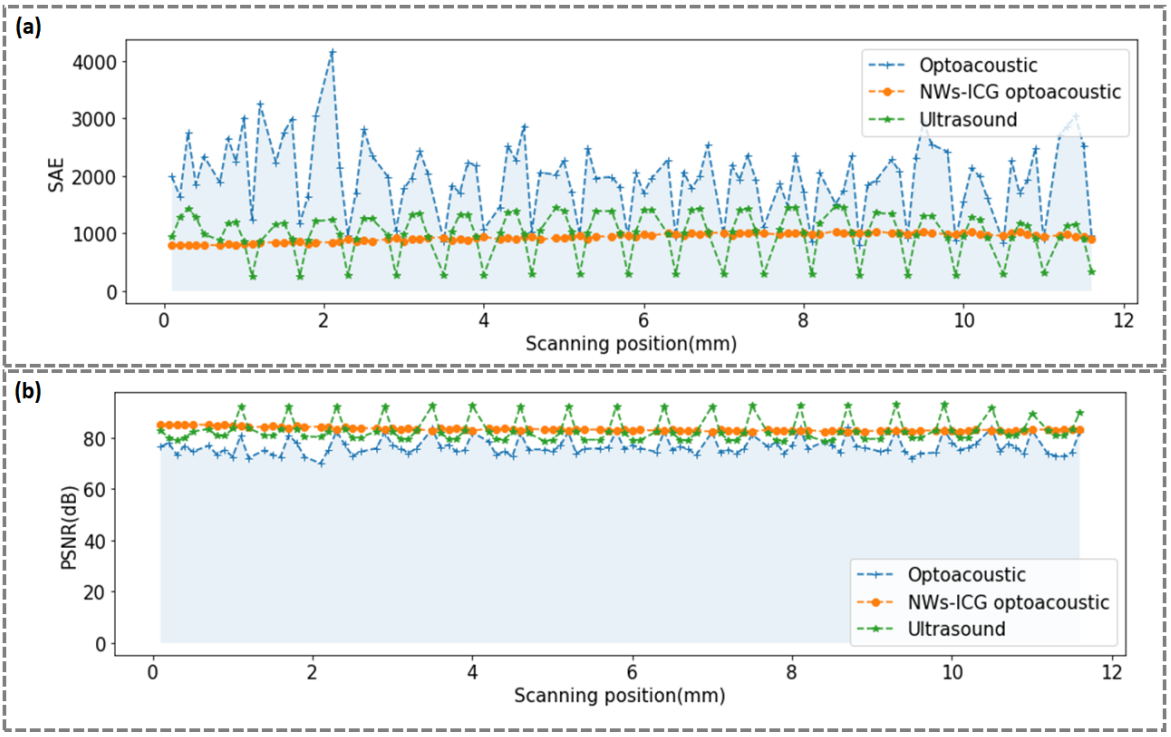


Figure 6. The PSNR and SAE (GEN, GT) evaluation in one of the testing tumors. (a-b) The graph between the PSNR and SAE (GEN, GT) values vs. scanning positions for all generated OPUS, NWs-ICG optoacoustic, and optoacoustic (λ_{ex} = 800 nm) images, respectively.

Table. 1 Average quantitative metrics of optoacoustic (λ_{ex} = 800 nm), NWs-ICG optoacoustic, and ultrasound images generated by the proposed deep learning model.

Average quantitative metrics	Optoacoustic (λ_{ex} = 800 nm)	NWs-ICG optoacoustic	Ultrasound
PSNR (dB)	76.53	83.75	80.44
SAE (GEN, GT)	**1,706.12**	858.54	**1,265.87**
SAE $(Input_1, GT)$	6,812.92	406.59	6,695.71
SAE $(Input_2, GT)$	5,294.94	**284.02**	4,902.67

4 | CONCLUSION

This work demonstrates a deep learning technique based on recurrent and convolution neural networks for generating sequential NWs-ICG optoacoustic (multispectral unmixing), ultrasound, and optoacoustic images. It has shown robust and promising performance in the accurate reconstruction of the sequential images for all modalities, according to the quantitative evaluation of model performance using the PSNR and SAE for all scanning positions of the generated images (reconstructed by the deep learning model) and ground truth (acquired by mechanical scanning). The architecture of our model is versatile since it can promisingly generate sequential cross-sectional images of three modalities from the commercial MSOT system. Using our deep learning can substantially reduce acquisition time. However, all the training data were acquired from *ex vivo* tissues completely fixed in agarose. Model performance with images acquired *in vivo* may be affected by cardiac and respiratory motion. In the future, we will explore the possibility of optimizing and applying the model to generate sequential images of *in vivo* samples with motion artifacts.

ACKNOWLEDGMENTS

We thank Amy Porter, Investigative Histopathology Laboratory, Michigan State University, for providing the H&E slides.

FINANCIAL DISCLOSURE

This work was supported by the National Science Foundation (NSF) (grant numbers 1808436, 1918074, and 2237142-CAREER), the Department of Energy (DOE) (grant number 234402), and partially supported by the National Research Council of Thailand under Grant FRB650025/0258, Grant NRCT.MHESRI/505/2563-65, Grant RE-KRIS-FF65-14/FF65-38, and Grant RE-KRIS-FF66-63/FF66-64.

CONFLICT OF INTEREST

The authors declare no conflicts of interest related to this article.

DATA AVAILABILITY STATEMENT

The data supporting this study's findings and the CNN trained in this study are available from the corresponding author upon reasonable request.

ORCID

Aniwat Juhong	https://orcid.org/0000-0002-9115-9767
Bo Li	https://orcid.org/0000-0002-4973-1969
Chia-wei Yang	https://orcid.org/0000-0002-0950-5510
Cheng-you Yao	https://orcid.org/0000-0002-6045-8676
Yifan Liu	https://orcid.org/0000-0003-2887-7704
Dalen W. Agnew	https://orcid.org/0000-0001-8538-1348
Yu Leo Lei	https://orcid.org/0000-0002-9868-9824
Gary Luker	https://orcid.org/0000-0001-6832-2581
Harvey Bumpers	https://orcid.org/0000-0001-6832-2581
Xuefei Huang	https://orcid.org/0000-0001-6165-9261
Wibool Piyawattametha	https://orcid.org/0000-0002-2228-8485
Zhen Qiu	https://orcid.org/0000-0001-8790-8481

REFERENCES

1. V. Ntziachristos and D. Razansky, "Molecular imaging by means of multispectral optoacoustic tomography (MSOT)," Chemical reviews **110**, 2783-2794 (2010).
2. L. V. Wang and S. Hu, "Photoacoustic tomography: in vivo imaging from organelles to organs," science **335**, 1458-1462 (2012).
3. A. Buehler, M. Kacprowicz, A. Taruttis, and V. Ntziachristos, "Real-time handheld multispectral optoacoustic imaging," Optics letters **38**, 1404-1406 (2013).
4. A. Dima and V. Ntziachristos, "In-vivo handheld optoacoustic tomography of the human thyroid," Photoacoustics **4**, 65-69 (2016).
5. A. C. Tam, "Applications of photoacoustic sensing techniques," Reviews of Modern Physics **58**, 381 (1986).
6. D. Razansky, M. Distel, C. Vinegoni, R. Ma, N. Perrimon, R. W. Köster, and V. Ntziachristos, "Multispectral opto-acoustic tomography of deep-seated fluorescent proteins in vivo," Nature photonics **3**, 412-417 (2009).
7. S. Tzoumas, N. C. Deliolanis, S. Morscher, and V. Ntziachristos, "Unmixing molecular agents from absorbing tissue in multispectral optoacoustic tomography," IEEE transactions on medical imaging **33**, 48-60 (2013).
8. G. Diot, S. Metz, A. Noske, E. Liapis, B. Schroeder, S. V. Ovsepian, R. Meier, E. Rummeny, and V. Ntziachristos, "Multispectral optoacoustic tomography (MSOT) of human breast cancer," Clinical Cancer Research **23**, 6912-6922 (2017).
9. I. Quiros-Gonzalez, M. R. Tomaszewski, S. J. Aitken, L. Ansel-Bollepalli, L.-A. McDuffus, M. Gill, L. Hacker, J. Brunker, and S. E. Bohndiek, "Optoacoustics delineates murine breast cancer models displaying angiogenesis and vascular mimicry," British journal of cancer **118**, 1098-1106 (2018).
10. A. Ron, X. L. Deán-Ben, S. Gottschalk, and D. Razansky, "Volumetric optoacoustic imaging unveils high-resolution patterns

of acute and cyclic hypoxia in a murine model of breast cancer," Cancer research **79**, 4767-4775 (2019).
11. A. Taruttis, G. M. van Dam, and V. Ntziachristos, "Mesoscopic and macroscopic optoacoustic imaging of cancer," Cancer research **75**, 1548-1559 (2015).
12. M. R. Tomaszewski, M. Gehrung, J. Joseph, I. Quiros-Gonzalez, J. A. Disselhorst, and S. E. Bohndiek, "Oxygen-enhanced and dynamic contrast-enhanced optoacoustic tomography provide surrogate biomarkers of tumor vascular function, hypoxia, and necrosis," Cancer research **78**, 5980-5991 (2018).
13. A. P. Regensburger, L. M. Fonteyne, J. Jüngert, A. L. Wagner, T. Gerhalter, A. M. Nagel, R. Heiss, F. Flenkenthaler, M. Qurashi, and M. F. Neurath, "Detection of collagens by multispectral optoacoustic tomography as an imaging biomarker for Duchenne muscular dystrophy," Nature medicine **25**, 1905-1915 (2019).
14. W. Song, Z. Tang, D. Zhang, N. Burton, W. Driessen, and X. Chen, "Comprehensive studies of pharmacokinetics and biodistribution of indocyanine green and liposomal indocyanine green by multispectral optoacoustic tomography," RSC advances **5**, 3807-3813 (2015).
15. T. Anani, A. Brannen, P. Panizzi, E. C. Duin, and A. E. David, "Quantitative, real-time in vivo tracking of magnetic nanoparticles using multispectral optoacoustic tomography (MSOT) imaging," Journal of pharmaceutical and biomedical analysis **178**, 112951 (2020).
16. M. K. Gurka, D. Pender, P. Chuong, B. L. Fouts, A. Sobelov, M. W. McNally, M. Mezera, S. Y. Woo, and L. R. McNally, "Identification of pancreatic tumors in vivo with ligand-targeted, pH responsive mesoporous silica nanoparticles by multispectral optoacoustic tomography," Journal of controlled release **231**, 60-67 (2016).
17. D. Li, G. Zhang, W. Xu, J. Wang, Y. Wang, L. Qiu, J. Ding, and X. Yang, "Investigating the effect of chemical structure of semiconducting polymer nanoparticle on photothermal therapy and photoacoustic imaging," Theranostics **7**, 4029 (2017).
18. S. Wang, L. Zhang, J. Zhao, M. He, Y. Huang, and S. Zhao, "A tumor microenvironment–induced absorption red-shifted polymer nanoparticle for simultaneously activated photoacoustic imaging and photothermal therapy," Science Advances **7**, eabe3588 (2021).
19. J. Gröhl, M. Schellenberg, K. Dreher, N. Holzwarth, M. D. Tizabi, A. Seitel, and L. Maier-Hein, "Semantic segmentation of multispectral photoacoustic images using deep learning," arXiv preprint arXiv:2105.09624 (2021).
20. A. Y. Yuan, Y. Gao, L. Peng, L. Zhou, J. Liu, S. Zhu, and W. Song, "Hybrid deep learning network for vascular segmentation in photoacoustic imaging," Biomedical Optics Express **11**, 6445-6457 (2020).
21. G. P. Luke, K. Hoffer-Hawlik, A. C. Van Namen, and R. Shang, "O-Net: a convolutional neural network for quantitative photoacoustic image segmentation and oximetry," arXiv preprint arXiv:1911.01935 (2019).
22. H. Lan, D. Jiang, C. Yang, and F. Gao, "Y-Net: a hybrid deep learning reconstruction framework for photoacoustic imaging in vivo," arXiv preprint arXiv:1908.00975 (2019).
23. J. Zhang, B. Chen, M. Zhou, H. Lan, and F. Gao, "Photoacoustic image classification and segmentation of breast cancer: a feasibility study," IEEE Access **7**, 5457-5466 (2018).
24. T. Chen, T. Lu, S. Song, S. Miao, F. Gao, and J. Li, "A deep learning method based on U-Net for quantitative photoacoustic imaging," in *Photons Plus Ultrasound: Imaging and Sensing 2020*, (International Society for Optics and Photonics, 2020), 112403V.
25. C. Bench, A. Hauptmann, and B. T. Cox, "Toward accurate quantitative photoacoustic imaging: learning vascular blood oxygen saturation in three dimensions," Journal of Biomedical Optics **25**, 085003 (2020).
26. C. Yang, H. Lan, H. Zhong, and F. Gao, "Quantitative photoacoustic blood oxygenation imaging using deep residual and recurrent neural network," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (IEEE, 2019), 741-744.
27. J. Gröhl, T. Kirchner, T. Adler, and L. Maier-Hein, "Estimation of blood oxygenation with learned spectral decoloring for quantitative photoacoustic imaging (LSD-qPAI)," arXiv preprint arXiv:1902.05839 (2019).
28. C. Cai, K. Deng, C. Ma, and J. Luo, "End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging," Optics letters **43**, 2752-2755 (2018).
29. D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," IEEE transactions on medical imaging **37**, 1464-1477 (2018).
30. N. Davoudi, X. L. Deán-Ben, and D. Razansky, "Deep learning optoacoustic tomography with sparse data," Nature Machine Intelligence **1**, 453-460 (2019).
31. A. Hariri, K. Alipour, Y. Mantri, J. P. Schulze, and J. V. Jokerst, "Deep learning improves contrast in low-fluence photoacoustic imaging," Biomedical optics express **11**, 3360-3373 (2020).
32. T. Lu, T. Chen, F. Gao, B. Sun, V. Ntziachristos, and J. Li, "LV-GAN: A deep learning approach for limited-view optoacoustic imaging based on hybrid datasets," Journal of biophotonics **14**, e202000325 (2021).
33. K. Sivasubramanian and L. Xing, "Deep learning for image processing and reconstruction to enhance led-based photoacoustic imaging," LED-Based Photoacoustic Imaging: From Bench to Bedside, 203-241 (2020).
34. B. Lafci, E. Merčep, S. Morscher, X. L. Deán-Ben, and D. Razansky, "Deep learning for automatic segmentation of hybrid optoacoustic ultrasound (OPUS) images," IEEE transactions on ultrasonics, ferroelectrics, and frequency control **68**, 688-696 (2020).
35. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), 234-241.
36. M. Aydın, B. Kiraz, F. Eren, Y. Uysallı, B. Morova, S. C. Ozcan, C. Acilan, and A. Kiraz, "A Deep Learning Model for Automated Segmentation of Fluorescence Cell images," in *Journal of Physics: Conference Series*, (IOP Publishing, 2022), 012003.
37. K. de Haan, H. Ceylan Koydemir, Y. Rivenson, D. Tseng, E. Van Dyne, L. Bakic, D. Karinca, K. Liang, M. Ilango, and E. Gumustekin, "Automated screening of sickle cells using a smartphone-based microscope and deep learning," NPJ digital medicine **3**, 76 (2020).
38. N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," Neural networks **121**, 74-87 (2020).
39. N. S. Punn and S. Agarwal, "Modality specific U-Net variants for biomedical image segmentation: a survey," Artificial Intelligence Review **55**, 5845-5889 (2022).
40. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016), 2818-2826.
41. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation **9**, 1735-1780 (1997).
42. M. Xu and L. V. Wang, "Universal back-projection algorithm for photoacoustic computed tomography," Physical Review E **71**, 016706 (2005).
43. N. Stergiou, N. Gaidzik, A.-S. Heimes, S. Dietzen, P. Besenius, J. Jäkel, W. Brenner, M. Schmidt, H. Kunz, and E. Schmitt, "Reduced Breast Tumor Growth after Immunization with a Tumor-Restricted MUC1 Glycopeptide Conjugated to Tetanus ToxoidImmunization against Tumor-Restricted MUC1 in Breast Cancer," Cancer Immunology Research **7**, 113-122 (2019).

44. C.-W. Yang, K. Liu, C.-Y. Yao, B. Li, A. Juhong, Z. Qiu, and X. Huang, "Indocyanine Green-Conjugated Superparamagnetic Iron Oxide Nanoworm for Multimodality Breast Cancer Imaging," ACS Applied Nano Materials **5**, 18912-18920 (2022).
45. K. Greish, "Enhanced permeability and retention (EPR) effect for anticancer nanomedicine drug targeting," Cancer nanotechnology: Methods and protocols, 25-37 (2010).
46. N. Keshava and J. F. Mustard, "Spectral unmixing," IEEE signal processing magazine **19**, 44-57 (2002).
47. S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015), 802-810.
48. A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," IEEE signal processing magazine **35**, 53-65 (2018).
49. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).